Hypothesis verification with histogram of compositions improves object detection of hierarchical models

Domen Tabernik[1], **Matej Kristan**[1], **Marko Boben**[1], **Aleš Leonardis**[1,2]

[1]*Faculty of Computer and Information Science, University of Ljubljana*
[2]*CN-CR Centre, School of Computer Science, University of Birmingham*
{domen.tabernik},{matej.kristan},{marko.boben},{ales.leonardis}@*fri.uni-lj.si*

Abstract

This paper focuses on applying and evaluating the additional hypothesis verification step for the detections of learnt-hierarchy-of-parts (LHOP) method. The applied method reduces the problem of false positives that are a common problem of hierarchical methods specifically in highly textured or cluttered images. We use a Histogram of Compositions (HoC) with a Support Vector Machine in hypothesis verification step. Using HoC descriptor ensures that the additional computation cost is as minimal as possible since HoC descriptor shares the LHOP tree structure. We evaluate the method on the ETHZ Shape Classes dataset and show that our method outperforms the original baseline LHOP method by around 5 percent.

1 Introduction

In the field of computer vision many different approaches have been proposed for solving the problem of object detection. One of the most effective and popular approach is using sliding windows. This approach slides a window across the whole image to create a dense set of regions from which features, such as HOG [2] or SIFT [11], are extracted and classified into a known object category. The most successful method based on this approach has been proposed by Felzenszwalb *etal*. using the discriminative deformable parts models [5]. This model is based on HOG features and utilizes a constellation-like approach to describe each object representation as two-level filter. While this method produces state-of-the-art results on PASCAL datasets [4], its main problem is rooted in the sliding windows approach which requires extensive computational resources to verify hundreds of thousands of regions per single image.

To avoid sliding windows many hierarchical models [9, 10, 14] have proposed to build detections from bottom-up approach in a layer-by-layer manner. This approach ensures that only simple features of lower layers are extracted over the whole image while at higher layer more complex shape representations of object categories are being used only from the most important parts of the image. Additionally, the hierarchical methods allow for sharing of parts [7] within their hierarchical structure thus allowing for more compact representation of object categories. Across multiple categories this positively affects detection speed. When a common shape is found in the lower layers of the hierarchies the detected shape can be reused in detection of multiple categories in higher layers that share this common shape. Compared to the current state-of-the-art methods the Achilles heel of such hierarchical methods is poorer performance. The problem is rooted in the bottom-up object inference where the inference process does find the correct object in the image, but at the same time it also *hallucinates* objects in highly textured or highly cluttered images. This produces many false positive detections that result in poorer performance as they cannot be easily removed by simple thresholding. To address the problem of false positive detections we can utilize the hypothesis verification step introduced by [13].

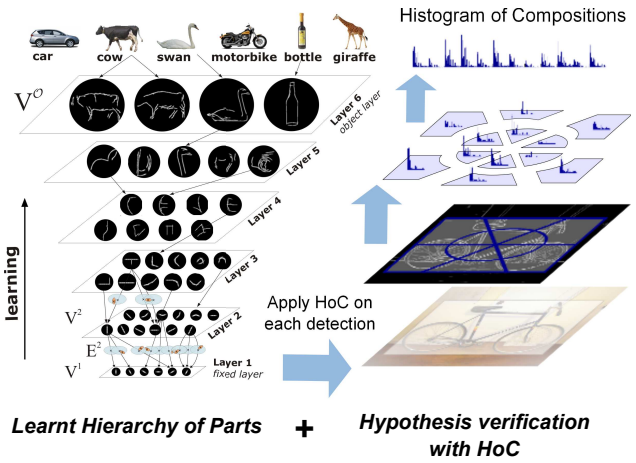

Figure 1: Applying the hypothesis verification to the learnt-hierarchy-of-parts (LHOP) detections. As hypothesis verification we utilize Histogram of Compositions that uses the same hierarchical model as LHOP method.

In our paper we apply hypothesis verification step to the hierarchical method (see, Figure 1) and as our contribution we show the performance of using this method on the *ETHZ Shape Classes [6]* dataset. The hypothesis veri-

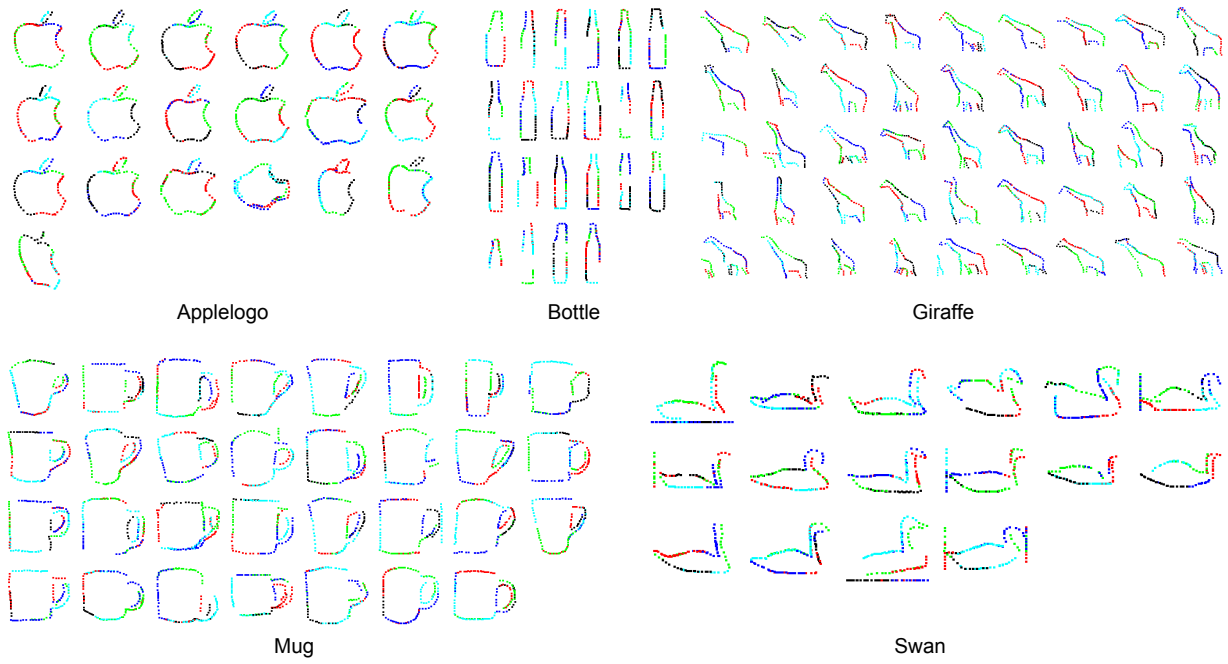

Figure 2: Models from object layer of LHOP method for five different categories of the *ETHZ Shape Classes* dataset.

fication step is applied after hierarchical method produces detections. We compute descriptor from the regions of the hypothetical object and use Support Vector Machine to verify the validity of this region. As a descriptor we use Histogram of Compositions (HoC) [12]. This descriptor allows us to have a fully integrated framework since HoC descriptor is composed from the same hierarchical tree as detected objects. The benefit of HoC is that it uses only lower layer shapes which have more discriminative information than upper layer object shapes from detections. We show that by using this discriminative information we are able to eliminate false positive detections and improve final performance of the hierarchical method.

The remainder of the paper is structured as follows. In Section 2 we provide formal description of using hypothesis verification with HoC descriptor, in Section 3 we present results and we conclude with the discussion in Section 4.

2 Hypothesis verification

We utilize the learnt-hierarchy-of-parts model (LHOP) [9] to produce hypothetical object detection. The same model is also used by the HoC descriptor that we apply in the hypothesis verification step. This allows us to use the same hierarchical tree for obtaining detections and for producing HoC descriptor. Although we use the same hierarchical model that produced false positive detections we construct HoC descriptor from simpler lower layer shapes that contain more discriminative information then upper layer shapes used for object detection. The higher layer objects are composed from lower layer shapes with response value only above certain threshold. This eliminates construction of noisy high level shapes with low response values but it also removes some of the discriminative information that is present in the lower layer parts. We can now bring this discriminative information back into the object detection by using them in our hypothesis verification step.

2.1 Detection with learnt hierarchy of parts

We now provide a simple notation for LHOP model, while we refer reader to [9] for further details. In the following we will denote the library of hierarchical parts trained for up to L layers as a set of N compositions $\mathcal{L} = \{P_i^l\}_{i=1:N}$, where P_i^l is an identifier of i-th composition and belongs to the l-th layer of the library. At the last layer L, in our case 6 layers were enough to capture complexity in categories, each composition directly identifies one trained category, i.e. for each category we have only one corresponding composition on the L-th layer. Applying the library $\mathcal{L}$ on a given image $\mathcal{I}$, the algorithm of hierarchical model infers a set of K detected parts, $\mathcal{C}(\mathcal{I}, \mathcal{L})$,

$$\mathcal{C}(\mathcal{I}, \mathcal{L}) = \{\pi_k^l\}_{k=1:K},$$

where the k-th detected part on the l-th layer $\pi_k^l = [P_k, \mathbf{c}_{\pi_k}, \lambda_k]$ is defined by its library identifier P_k^l, its location $\mathbf{c}_{\pi_k}$ in the image and its detection score λ_k . All the inferred parts from the last layer L directly correspond to detected objects in the image:

$$\mathcal{D}(\mathcal{I}, \mathcal{L}) = \{\pi_j^L\}_{j=1:J},$$

where $\mathcal{D}(\mathcal{I}, \mathcal{L})$ is a set of J detected objects in the image $\mathcal{I}$ processed with the library $\mathcal{L}$. While each detected

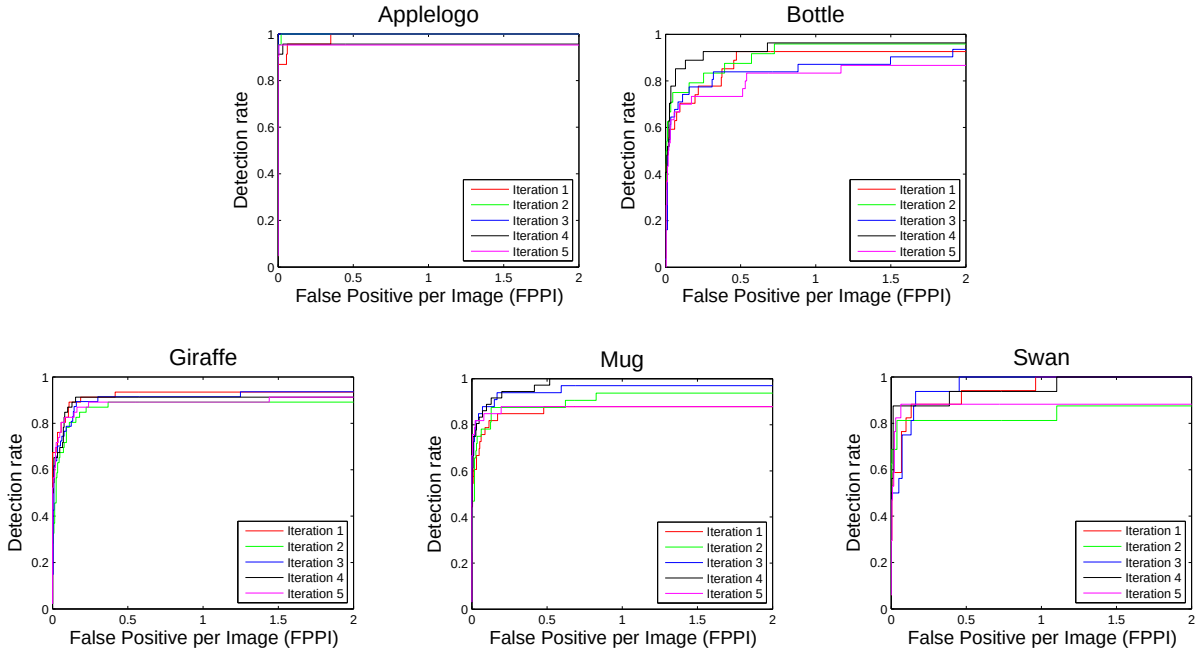

Figure 3: Detection rate over different FPPI rates (false positives per image) for each category from the *ETHZ Shape Classes* dataset. We ran five iterations and sampled examples randomly in each iteration to obtain different training/testing splits.

object is defined the same as detected part at L-th layer $\pi_j^L = [P_j^L, \mathbf{c}_{\pi_j}, \lambda_j]$, we can also add a category information since a library identifier P_j^L from the L-th layer always directly matches to one learning category. We can also obtain a bounding box location of detected object simply by tracing down sub-parts of detected part π_j^L to the first layer. Minimal and maximal locations of all traced sub-parts define a bounding box of detected image. We can therefore define a set of detected objects from a given image $\mathcal{I}$ as:

$$\mathcal{D}(\mathcal{I}, \mathcal{L}) = \{(\pi_j^L, c_j, r_j)\}_{j=1:J},$$

where c_j is detected category and $r_j = (x, y, w, h)$ is a detection bounding box.

2.2 Hypothesis verification with HoC descriptor

For each detected object $\mathcal{D}(\mathcal{I}, \mathcal{L})$ we can obtain category information c_j and detected bounding box r_j. Within a bounding box r_j we calculate a HoC descriptor $\mathcal{H}_j$ from detected parts of second and third layer. In [12] the authors use library pre-trained on a general set of images but since the HoC descriptor incorporates the same LHOP model as we use for the detection we can easily compute descriptor on the same library of compositions. This also eliminates the time required to reprocess the image with different library. All computed descriptors $\mathcal{H}_j$ are then filtered by the Support Vector Machine with a category model c_j. As the final step we perform a non-maximum suppression using a greedy approach.

3 Experiments and results

We evaluated the method on the *ETHZ Shape Classes [6]* dataset and compared it to the results of the original baseline method of LHOP from [8]. We repeated the experiments five times, independently for each category, using the following procedure: half of the images that contained the selected category where randomly chosen for the training, while the other half were used for the testing only. All other images that did not contain category objects where used only for testing. Selected category training images were used to train LHOP library $\mathcal{L}$ for up to 6 layers (see, Figure 2). To train SVM for the selected category we extracted object regions from training images and computed HoC descriptor $\mathcal{H}_j$ for positive examples. The training objects were scaled to 120 pixel wide regions. We collected negative examples by running LHOP with trained library $\mathcal{L}$ on training images to produce HoC descriptor from hypothetical detections. The training images for negatives were first resized by a factor of 1.2 and then scaled by a factor of $\sqrt{2}$ to produce around 4-7 scales per image (the same scaling procedure was followed during the testing stage). As hard-negatives we used all detections that had PASCAL intersection criteria with the ground truth less then 0.3 and as additional positive examples we used all detections with intersection criteria of more then 0.7. Due to a small number of training images from the ETHZ dataset we also sampled Caltech-101 [3] dataset for additional hard-negative examples. On average we used around 40 000 negative examples

	Apple logo	Bottle	Giraffe	Mug	Swan	*Average*
LHOP only [8] (baseline)	88.2 (3.4)	**87.6 (1.5)**	83.5 (1.1)	86.1 (2.0)	80.0 (3.5)	*85.1*
LHOP + HoC verification (our)	**98.2 (2.5)**	84.5 (7.1)	**90.5 (1.2)**	**89.7 (4.2)**	**89.0 (5.2)**	***90.4***

Table 1: Evaluation result on *ETHZ Shape Classes* with reported detection-rate (%) at 0.4 FPPI averaged over five iterations (standard deviation values are shown in parentheses)

and 500 positive examples. Support Vector Machine was implemented using LIBSVM [1] with an RBF kernel using chi-squared distance function (RBF-$\mathcal{X}^2$).

3.1 Results

Summary of detection rates are reported in Table 1 with additional detection-rate versus FPPI curves for all categories across different iterations in Figure 3. We compared our results to the baseline results of [8]. Their method used only LHOP models from object layers with simple threshold filtering to produce the best possible results. Based on the reported results we notice considerable improvements across all five categories. Hypothesis verification outperformed baseline method for all categories except for the bottle. The performance drop for this category was 3 percent but due to high standard deviation the difference is not significant. On average, our method performed by around 5 percent better then baseline LHOP method, producing detection rate of 90.4 percent versus 85.1 percent for LHOP at 0.4 false positives per image (FPPI).

4 Conclusion

In this paper we have applied additional hypothesis verification step to the hierarchical methods. Specifically we used learnt-hierarchy-of-parts (LHOP) model [9] for object detection and applied Histogram of Compositions [12] descriptor on detected hypothetical objects. Using discriminative information from HoC descriptor in Support Vector Machine we were able to eliminate many false positive detections that are commonly occurring in highly textured or cluttered images. We have demonstrated the performance on the *ETHZ Shape Classes [6]* dataset where we outperformed baseline LHOP method by 5 percent and achieved detection-rate of 90.4 percent at 0.4 false positives per image.

In future work we plan on evaluating this method on bigger datasets (e.g. PASCAL) and comparing it to other state-of-the-art methods. Additionally, we would also like to apply more texture based descriptors such as local-binary-pattern as the current descriptor is mostly shape oriented.

Acknowledgments. This work was supported in part by ARRS research program P2-0214 and ARRS research projects J2-4284, J2-3607 and J2-2221.

References

[1] Chih Chung Chang and Chih Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[3] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. In *IEEE Transactions on Pattern Recognition and Machine Intelligence*. IEEE Trans., 2004.

[4] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://www.cs.brown.edu/ pff/latent-release4/.

[5] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[6] Vittorio Ferrari, Tinne Tuytelaars, and Luc Van Gool. Object detection by contour segment networks. In *Proceeding of the European Conference on Computer Vision*, volume 3953 of *LNCS*, pages 14–28. Elsevier, June 2006.

[7] Sanja Fidler, M. Boben, and Aleš Leonardis. Evaluating multi-class learning strategies in a generative hierarchical framework for object detection. In *Neural Information Processing Systems*, 2009.

[8] Sanja Fidler, Marko Boben, and Aleš Leonardis. A coarse-to-fine taxonomy of constellations for fast multi-class object detection. In *ICCV*, pages 687–700, Berlin, Heidelberg, 2010. Springer-Verlag.

[9] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*. IEEE Computer Society, 2007.

[10] Iasonas Kokkinos and Alan Yuille. Inference and learning with hierarchical shape models. *Int. J. Comput. Vision*, 93(2):201–225, June 2011.

[11] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[12] D. Tabernik, M. Kristan, M. Boben, and A. Leonardis. Learning statistically relevant edge structure improves low-level visual descriptors. In *International Conference on Pattern Recognition*, 2012.

[13] Domen Tabernik, Luka Čehovin, Matej Kristan, Marko Boben, and Aleš Leonardis. A web-service for object detection using hierarchical models. In *The 9th International Conference on Computer Vision Systems*, 2013.

[14] Long Zhu, Yuanhao Chen, A. Torralba, W. Freeman, and A. Yuille. Part and appearance sharing: Recursive compositional models for multi-view. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1919 –1926, june 2010.